Evaluation of the Moral Permissibility of Action Plans*

Felix Lindner, Robert Mattmüller and Bernhard Nebel
University of Freiburg, Germany

Abstract

Research in classical planning so far has been mainly concerned with generating a satisficing or an optimal plan. However, if such systems are used to make decisions that are relevant to humans, one should also consider the ethical consequences generated plans can have. Traditionally, ethical principles are formulated in an action-based manner, allowing to judge the execution of one action. We show how such a judgment can be generalized to plans. Further, we study the computational complexity of making ethical judgment about plans.

1 Introduction

With the advent of autonomous machines that drive on the streets or act as household robots, it has been argued that we need to add an ethical dimension to such machines leading to the development of the research area *machine ethics* (Anderson, Anderson, & Armen, 2005; Anderson & Anderson, 2011). One important question is how we can align the behavior of autonomous machines with the moral judgment of humans. In this context, most often the question is whether a particular action is morally obligatory, permissible or impermissible, given a particular ethical principle (Driver, 2006). Judging one action is, of course, important. However, automated planning systems (Ghallab, Nau, & Traverso, 2016) are faced with the problem of making a huge number of decisions about including actions into a plan. And it does not necessarily make sense to analyze the ethical contents of each such decision in isolation, but it may be necessary to take an ethical perspective on an entire plan (and perhaps alternative plans). This is especially true if it should be possible for an autonomous agent to cause harm first and repair it later. As an example, consider utilitarian reasoning: if every action in a plan were judged in isolation, one would not be allowed to perform an action that temporarily decreases the utility, even if this action is a necessary prerequisite for later earning a lot of utility in a globally

*This is an extended and revised version of the paper *Moral Permissibility of Action Plans* published in the *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*. It will be published in *Artificial Intelligence*.

optimal final reachable state. Judging a plan as a whole allows considering this early investment for the sake of a later benefit as permissible from a utilitarian perspective.

In this article, we address these problems in the setting of classical AI planning. First, we will look at what kind of additional information we need in order to be able to make moral judgments in the context of different ethical theories. Secondly, we will propose methods to judge the ethical acceptability of a plan. We will test the proposed notions using examples from the literature on moral dilemmas. Thereby, we do not limit ourselves to one particular ethical principle, but will consider a number of different principles that have the potential to be treated computationally, similar to the HERA (Lindner, Bentzen, & Nebel, 2017) approach. Third, we will analyze the computational complexity of assessing the moral permissibility of a plan. Note that we do not intend to develop a new architecture for ethical reasoning agents. So, if one wants to use our approach it needs to be intergated in such an architecture.

The remainder of the paper is structured as follows: In the next section, we introduce different ethical principles that have been discussed in the literature. Then, the planning formalism we will use throughout the paper will be specified. This is basically a propositional planning formalism extended by variables with non-binary domains, exogenous events, and moral valuations of actions and consequences. We then formalize the notions of causation and means to an end in the framework of our planning formalism. Based on that, we can then formalize different ethical principles, which we will use to analyze the computational problem of ethically validating a given plan. Finally, we sketch related work and conclude.

2 Ethical Principles

Ethics is a subfield of practical philosophy and itself a broad area of research. At its core, ethics is concerned with the question of how agents ought to act. Specifically, normative ethics investigates ethical principles of acting morally correct. Traditionally, three classes of ethical principles can be distinguished: deontology, consequentialism, and virtue ethics.

Virtue ethics goes back to Aristotle. The ethical principle virtue ethics asks agents to follow is to live a good life by realizing virtues, such as courage, truthfulness, and modesty. Virtue ethics thus involves very bold concepts hard to formalize, and we will not deal with virtue ethics in this paper (but see (Govindarajulu2019, Bringsjord, Ghosh, & Sarathy, 2019) for a recent attempt to formalize virtue ethics).

The second class of ethical principles is called consequentialism. According to consequentialist ethics, the moral permissibility of an act is determined by its consequences. The most well-known member of this class is the *utilitarian principle* advocated by Jeremy Bentham and John Stuart Mill in the 18th and 19th century. This principle says that an agent ought to perform the act amongst the available alternatives that leads to the maximum utility, where utility is

measured in terms of pain and happiness. Sometimes it is also referred to as "the greatest happiness principle". We will formalize a version of the utilitarian principle in Subsect. 5.2.1.

Do-no-harm principles like the ones formulated in Asimov's first law of robotics are also consequentialist in nature: They asks agents to not cause harm and to not let harm happen. While the utilitarian principle will allow harm if it leads to overall maximum utility, the do-no-harm principle will be more rigorous. Moreover, for do-no-harm principles the distinction between *doing* and *allowing* is relevant: The first do-no-harm principle we will look at forbids actively causing harm but allows letting harm happen (see Subsect. 5.2.2). A more restrictive version of the do-no-harm principle is *Asimov's first law of robotics* additionally forbidding robots to letting harm happen through inaction when harm could be avoided by acting (see Subsect. 5.2.3). As another variant of do-no-harm, we will introduce the *do-no-instrumental-harm principle*. This principle allows to distinguish harm caused as a means to an agent's goal, and harm as a mere side effect (see Subsect. 5.2.4). For example, if a doctor gives medication to a patient to lower the patient's pain while knowing that the medication will slightly negatively affect the lever, the do-no-harm principle will forbid the medication, while the do-no-instrumental-harm principle will allow it.

Because the do-no-instrumental-harm principle, according to our formulation, also takes the agent's goal into account rather than consequences only, this principle is at the interface to the third class of ethical principles, i.e., *deontology*. Deontological ethics claims that moral permissibility of an act cannot be determined by the act's consequences alone. We will formulate two variants of deontology in Subsect. 5.1: First, we will consider the case were an act has an intrinsic moral value, and this value is all that counts to judge the act's moral permissibility. This moral value could, for instance, stem from a list like the ten commandments—given rules to perform or refrain from performing that act. We call this principle *act-based deontological principle*. Another type of deontology focuses on the agent's goals. This goes back to Kant who writes in *Groundwork of the Metaphysic of Morals* that only a good will can be morally good. Thus, an act permitted by act-based deontology may yet be forbidden when the agent's intentions a not morally good. Thus, we introduce *goal-based deontological principle*, which focuses on the agent's goal. We will, however, not make any attempt to formalize Kant's categorical imperative (but see (Lindner & Bentzen, 2018)).

Finally, we consider the *principle of double effect*. This principle has its origins in Catholic theology, cf., (Mangan, 1949), and has been applied, within philosophy, to trolley problems by Thomson (1985) as a reply to Foot's analysis of the problem of abortion (Foot, 1967). The principle of double effect is a mixture of deontological and consequentialist principles. It seeks to forbid intrinsically bad actions, bad intentions, and disproportionate consequences. Particularly, according to the principle of double effect, an act is permissible if the following five conditions hold (Mangan, 1949):

1. The action itself is morally good or neutral.

2. Some positive consequence is intended.
3. No negative consequence is intended.
4. No negative consequence is a means to the goal.
5. The positive consequences sufficiently outweigh the negative ones.

The first condition of the principle of double effect implements act-based deontology. Thus, actions are assumed to have an inherent moral value, which does not (necessarily) stem from the effect of an action. The second and third conditions take the intentions, or goals, of the agent into consideration: An agent may not have a bad consequence as a goal, but it should intend something good. The fourth condition is an implementation of the do-no-instrumental-harm principle: Morally bad consequences are permissible as side effects only. And finally, the fifth condition is a weaker version of utilitarianism: In our interpretation, this condition requires that all in all the effects of the action must yield positive utility (cf., (Bentzen, 2016)).

The principle of double effect has already been formalized within machine ethics, among others by Bentzen (2016), Govindarajulu and Bringsjord (2017), Pereira and Saptawijaya (2017), Hölldobler (2018). We contribute a formalization within classical planning in Subsect. 5.3.

3 Planning Formalism

We use a planning formalism based on SAS$^+$ (Bäckström & Nebel, 1995), extended with conditional effects (Rintanen, 2003) and exogenous events (Fox, Howey, & Long, 2005; Cresswell & Coddington, 2003). In order to permit to judge a plan for its ethical value, this basic formalism is further extended with means to specify utility values for actions and facts. Furthermore, we adopt a slightly non-standard execution semantics.

3.1 Language

A *planning task* is a tuple $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ consisting of the following components: $\mathcal{V}$ is a finite set of *state variables* v, each with an associated finite domain $\mathcal{D}_v$. A *fact* is a pair $\langle v, d \rangle$, where $v \in \mathcal{V}$ and $d \in \mathcal{D}_v$, also written as $v{=}d$ in conditions and $v{:=}d$ in effects. The set of all facts is denoted by F. We call a conjunction of facts $v_1{=}d_1 \wedge \cdots \wedge v_k{=}d_k$ *consistent* if it does not contain any two facts $v_i{=}d_i$ and $v_j{=}d_j$ such that $v_i = v_j$, but $d_i \neq d_j$. We call it a *complete conjunction*, or simply *complete* if it contains a conjunct $v{=}d$ for every variable $v \in \mathcal{V}$. Up to reordering and unnecessary repetitions of conjuncts, there is a unique complete conjunction of facts for every possible assignment of domain values to variables. Therefore, we will often identify those representations. A complete conjunction of facts s is also called a *state*, and S denotes the set of states of Π. The set A is a set of *actions*, where an action is a pair $a = \langle pre, eff \rangle$. The *precondition* pre is a conjunction of facts, and the *effect*

eff is a *conditional effect* in effect normal form (ENF) (Rintanen, 2003), i. e., a conjunction $\mathit{eff} = \mathit{eff}_1 \wedge \cdots \wedge \mathit{eff}_k$ of sub-effects eff_i of the form $\varphi_i \rhd v_i{:=}d_i$, where φ_i is a conjunction of facts, the *effect condition*, and where $v_i{:=}d_i$ is an atomic effect (a fact). Every atomic effect may occur at most once in *eff*. We furthermore assume that, whenever *eff* includes two conjuncts $\varphi_i \rhd v_i{:=}d_i$ and $\varphi_j \rhd v_j{:=}d_j$, and $v_i = v_j$, but $d_i \neq d_j$, then $\varphi_i \wedge \varphi_j$ is inconsistent, to rule out contradictory effects. If some φ_i is the trivial condition $\top$ (true), then the corresponding sub-effect is unconditional, and we write $v{:=}d$ instead of $\top \rhd v{:=}d$. The set of actions A is partitioned into a set A_{endo} of *endogenous actions* (controllable, planned for and executed by the planning agent) and a set A_{exo} of *exogenous actions* (uncontrollable by the planning agent, executed by the environment/nature). We assume that the set of endogenous actions always contains the *empty action* ϵ, which has an empty precondition and effect, and we assume that each exogenous action is associated with a set of discrete time points $t(a)$ at which it will be automatically applied, provided that its preconditions is satisfied. This is similar in spirit to *timed facts* (Cresswell & Coddington, 2003) that are made true exactly at their associated time point. The state $s_0 \in S$ is called the *initial state*, and the partial state $s_\star$ specifies the *goal* condition.

3.2 Semantics

An endogenous action $a = \langle \mathit{pre}, \mathit{eff} \rangle$ is applicable in state s iff $s \models \mathit{pre}$, i. e., the precondition *pre* is satisfied in s. For an exogenous action a to be applicable, we additionally require that s is the t-th state in the state sequence induced by the action sequence under consideration for some $t \in t(a)$. Let $\mathit{eff} = \bigwedge_{i=1}^{k}(\varphi_i \rhd v_i{:=}d_i)$ be an effect in ENF. Then the *change set* (Rintanen, 2003) of *eff* in s, symbolically $[\mathit{eff}]_s$, is the set of facts $\bigcup_{i=1}^{k}[\varphi_i \rhd v_i{:=}d_i]_s$, where $[\varphi \rhd v{:=}d]_s = \{v{=}d\}$ if $s \models \varphi$, and $\emptyset$, otherwise. A change set will never contain two contradicting effects. Now, applying an applicable action a to s yields the state s' that has a conjunct $v{=}d$ for each $v{=}d \in [\mathit{eff}]_s$, and the conjuncts from s for all variables v that are not mentioned in the change set $[\mathit{eff}]_s$. We write $s[a]$ for s'.

For exogenous actions, we assume an *urgent* semantics. More specifically, whenever an exogenous action a_{exo} is applicable and its application in the current state leads to a different successor state, its application is enforced. We furthermore assume that if two or more exogenous actions are applicable in the same state, they do not interfere, i. e., neither of them disables another one, nor do they have conflicting effects. Let s be a state. Then by $\Delta_{\text{exo}}(s)$ we refer to the unique state that is obtained from s by applying all applicable exogenous actions. Since exogenous actions that are applicable in the same time step do not interfere, $\Delta_{\text{exo}}(s)$ is well-defined and is obtained by the application of finitely many exogenous action occurrences. We give the following semantics to a sequence consisting of endogenous actions $\pi = \langle a_0, \ldots, a_{n-1} \rangle$: First we extend the sequence by empty actions if $n-1 < \max \bigcup_{a \in A_{\text{exo}}} t(a)$ until the highest time step of the exogenous actions equals $n-1$. Assume that the initial state s_0 is al-

ready closed under exogenous action application, i. e., that $\Delta_{\text{exo}}(s_0) = s_0$. Then, for $i = 0, \ldots, n-1$, the next state s_{i+1} is obtained by first applying action a_i to state s_i (assuming that it is applicable), followed by closing under exogenous actions. More formally, $s_{i+1} = \Delta(s_i, a_i) := \Delta_{\text{exo}}(s_i[a_i])$. If a_i is inapplicable in s_i for some $i = 0, \ldots, n-1$, then π is inapplicable in s_0.

A state s is a goal state if $s \models s_\star$. We call π a *plan* for Π if it is applicable in s_0 and if its last state s_n is a goal state, i.e., $s_n \models s_\star$.

3.3 Modified semantics for counterfactual reasoning

Below, we will propose a way to answer questions of the form: "What would have happened if we had followed plan π, but without action a being part of π?", or: "What would have happened if $v{:=}d$ had not been an effect of action a?" For that, we want to be able to trace plan π while leaving out a or $v{:=}d$. Unfortunately, with the semantics above, this would often simply mean that the modified plan is no longer applicable. To avoid this, we consider an alternative semantics here. Let $\pi' = \langle a_0, \ldots, a_{n-1} \rangle$ be a modified plan, possibly with some actions replaced by the empty action ϵ, or with some effects removed from actions. Let s_0 be the initial state. Then we define, for all $i = 0, ..., n-1$, that $s_{i+1} = \Delta(s_i, a_i)$, if a_i is applicable in s_i, and $s_{i+1} = \Delta(s_i, \epsilon)$, otherwise. In other words, if a_i is applicable in s_i, then we apply it, otherwise, we skip it. Notice that even if a_i remains applicable in s_i in π', the actual effects of a_i may differ from what happens when tracing the original plan π, since some effect conditions of a_i may be satisfied for π, but not for π', or the other way around.

Note that this non-standard semantics is equivalent to reformulating the planning domain slightly and executing plans under the standard semantics. In the reformulation, one would replace all preconditions with $\top$ and conjoin the original preconditions to the effect conditions. For example, an action *switch-light-on* $= \langle$*light-off*, *light-on*$\rangle$ can be replaced with *switch-light-on*$' = \langle \top,$ *light-off* $\rhd$ *light-on*$\rangle$.

3.4 Moral valuations of actions and consequences

Above, we defined the planning formalism we use. To define the possible *dynamics* of the system under consideration, this is sufficient. However, in order to formally capture and reason about the *ethical principles* outlined above, we also need to classify actions and facts with respect to their moral value as either morally bad, indifferent, or good. To that end, in the following, we assume that each planning task Π comes with a utility function u that maps endogenous actions and facts to utility values: $u\colon A_{\text{endo}} \cup F \to \mathbb{R}$. Notice that exogenous actions are not part of the moral evaluation, as they are not under control of the agent. Dealing with cases in which (the consequence of) an endogenous action ought to be judged in a different manner depending on its exogenous context is the responsibility of the ethical principles we will discuss below, not of the utility values assigned to endogenous actions.

We let u map to $\mathbb{R}$ instead of just $\{-1, 0, 1\}$ to allow for different degrees of how morally good or bad a fact may be. We need this in order to reasonably capture the utilitarian principle. In the case of actions a, on the other hand, we are only interested in the sign of $u(a) \in \{-1, 0, 1\}$. We call an action a or fact f *morally bad* if $u(a) < 0$ or $u(f) < 0$, respectively. Similarly, we call an action or fact *morally indifferent* or *morally good* if its utility value is zero or greater than zero, respectively. Notice that we explicitly do *not* require that moral values of actions and facts must be consistent in any particular sense. For instance, we do not require that an action must be classified as morally bad if one (or all) of its effects are morally bad. The rationale behind this choice is that, in terms of deontology, actions are good or bad *per se*, without regard to their actual effects. We leave enforcing such consistency to the modeler where this is desired, and emphasize that occasionally, such consistency may explicitly *not* desired.

When using a consequentialist view, we will judge the moral value of a plan based on the value of its final state, which is defined to be the sum over the utility values of all facts in the final state: $u(s) = \sum_{\{v=d \mid s \models v=d\}} u(v{=}d)$. Note that the actual value of $u(s)$ is only relevant to the utilitarian principle. All other ethical principles will base their decisions on other aspects, e.g., whether some morally bad fact in the final state is caused by the plan. If we want to consider also the utility value of intermediate states of a plan, one would need to propagate the relevant facts to the final state. This again would be something the modeler is responsible for.

4 Trolley Problems

To exemplify how the planning formalism SAS$^+$can be used to represent situations that involve ethical decisions, we present SAS$^+$models of two versions of the *trolley problem* (Foot, 1967), i.e., the classical trolley problem and the footbridge trolley problem.

The classical trolley problem is a thought experiment that asks the listener to imagine they were in the following situation: "A runaway trolley (i. e., tram) is about to run over and kill five people. If you, as a bystander, throw a switch then the trolley will turn onto a sidetrack, where it will kill only one person." So stated, the trolley problem has the following properties: First, the action possibilities are made explicit and there is no uncertainty about the respective outcomes. Second, there is no doubt that everyone affected deserves moral consideration, that is, the question of moral patiency is not an issue. Particularly, every death of a human on the track is equally bad. We acknowledge that reasoning about uncertainty (e.g., by taking probabilities into account) and about moral patiency (i.e., the determination of the moral value of actions and facts) are important problems for a full-fledged artificial moral agent. However, the scope of our work is to judge a plan under the assumption that these questions are already decided. Hence, trolley problems present a reasonable set of examples to demonstrate how our formalizations of various ethical principles in the

AI planning formalism behave.

Using SAS$^+$, the dynamics of the trolley problem can be modeled as a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, such that:

$$
\begin{aligned}
&\mathcal{V} = \{man, men, tram, lever\}\\
&A_{endo} = \{pull\}, A_{exo} = \{advance\}\\
&pull = \langle \top, lever{=}l \triangleright lever{:=}r \wedge lever{=}r \triangleright lever{:=}l \rangle\\
&advance = \langle \top, (tram{=}start \wedge lever{=}r) \triangleright tram{:=}r \wedge\\
&\qquad (tram{=}start \wedge lever{=}l) \triangleright tram{:=}l \wedge\\
&\qquad tram{=}r \triangleright men{:=}dead \wedge tram{=}l \triangleright man{:=}dead \rangle\\
&t(advance) = \{1, 2\}\\
&s_0 = man{=}alive \wedge men{=}alive \wedge tram{=}start \wedge lever{=}r\\
&s_\star = men{=}alive\\
&u(pull) = u(lever{=}l) = u(lever{=}r) = u(tram{=}start) =\\
&u(tram{=}l) = u(tram{=}r) = 0, u(man{=}alive) = 1,\\
&u(men{=}alive) = 5, u(man{=}dead) = -1,\\
&u(men{=}dead) = -5
\end{aligned}
$$

In this model, the variable *men* models the state of the five persons on the one track (*dead* or *alive*), and *man* models the state of the one person on the other track. The variable *tram* tracks the position of the tram (*start*, right track r, left track l), and the variable *lever* represents the state of the lever (left position l or right position r). There is one endogenous action *pull* available to the bystander. The action switches the state of the lever. The timed exogenous action *advance* changes the position of the tram at time points 1 and 2. Deaths have negative utility and survival facts have positive utility. All other facts and actions are considered morally neutral and thus have utility 0. Depending on the state of the lever, at time point 1, the tram will move from its start position either to the left track or to the right track. At time point 2, if it is on the left track, the tram will hit the one man, and if it is on the right track, it will hit the five men. So, if the bystander's goal was to save the five men, their only chance is to execute *pull* at time point 0.

The classical trolley problem is often contrasted with the *footbridge trolley problem*, which reads: "A trolley has gone out of control and now threatens to kill five people working on the track. The only way to save the five workers is to push a big man currently standing on the footbridge above the track. The big man will fall onto the track thereby stopping the tram. He will die, but the five other people will survive." Like the classical trolley problem, the footbridge trolley problem also involves a decision between one death and five deaths. For many people, however, the intuition about what is morally permissible to do turns out to be very different to that in the classical case. One explanation for this difference may be found in its differing causal dynamics, i.e., the man on the bridge is used as a means to save the life of the other five men. A SAS$^+$ model of

the footbridge trolley problem is given by the planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, such that:

$$\mathcal{V} = \{man, men\}, A_{endo} = \{push\}, A_{exo} = \{advance\}$$
$$push = \langle man{=}onBridge, man{:=}deadOnTrack \rangle$$
$$advance = \langle \top, man{=}onBridge \triangleright men{:=}dead \rangle$$
$$t(advance) = \{1\}$$
$$s_0 = man{=}onBridge \wedge men{=}alive, s_\star = men{=}alive$$
$$u(push) = -1, u(man{=}onBridge) = 1,$$
$$u(man{=}deadOnTrack) = -1, u(men{=}dead) = -5,$$
$$u(men{=}alive) = 5$$

The variable *man* represents the state of the big man on the footbridge (either *onBridge* or *deadOnTrack*), and the variable *men* represents the state of the five people on the track (either *dead* or *alive*). The endogenous action *push* is available to the decision-making agent, who reasons about whether or not to push the big man off the bridge. The timed exogenous action *advance* changes the state of the tram. Depending on whether or not the big man is on the track, the tram will stop at time point 1 due to its collision with the big man, or it will hit the other five men. We assume that pushing is inherently morally bad, that the fact that the big man is lying dead on the track is morally bad, that his surviving on the bridge is morally good, and that the death of the five men also is morally bad but their survival is morally good. In the modeled situation, the agent's goal is to save the five men.

We will refer to these two example domains during the next sections to analyze how our formalizations of ethical principles compare. When necessary, we will introduce further examples but we will not always provide full specifications of the planning tasks.

Finally, we want to stress that we do not claim that SAS$^+$is a perfect formalism for representing moral domains. Rather, our goal is to morally judge action plans that are generated by AI planning systems that make use of SAS$^+$planning task descriptions, e.g., FastDownward (Helmert, 2006) and its derivatives.

5 Formalization of Ethical Principles

The reasoning task of interest is to check possible plans for moral permissibility. To do so, we define moral permissibility of the ethical principles introduced above in a formal way. For deontology and utilitarianism, this turns out to be straightforward by using the utility assignment for actions and facts. For the remaining principles, we need some form of counterfactual reasoning in order to identify the cause of facts one wants to avoid.

5.1 Deontology

The definition of the deontological principle (Def. 1) requires that all actions in a plan are intrinsically morally good or neutral.

Definition 1 (Act-based Deontological Principle)**.** *Given a planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ *and an associated utility function* u*, a plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *for* Π *is* morally permissible according to the deontological principle *if and only if* $u(a_i) \geq 0$ *for all* $i = 0, \ldots, n-1$.

Note that Def. 1 only considers if the actions in the plan are individually morally bad or not. The deontological principles does not care about the degree of badness or goodness. Moreover, the moral value is assumed to be given by external means, e.g., by the modeler, or by some external process that evaluates the action against a set of moral rules. Consider the plans $\pi_1 = \langle pull \rangle$ for the classical trolley problem and $\pi_2 = \langle push \rangle$ for the footbridge trolley problem as modeled above. Plan π_1 does not contain any intrinsically bad action, whereas π_2 does. Therefore, according to the deontological principle, π_1 is morally permissible and π_2 is morally impermissible.

Some modifications of deontology propose that actions sould be evaluated based on the agent's goal. Definition 2 captures this view.

Definition 2 (Goal-based Deontological Principle)**.** *Given planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ *and an associated utility function* u*, a plan* π *for* Π *is* morally permissible according to the goal-based deontological principle *if and only if* $u(g) \geq 0$ *for all facts* g *that are part of the agent's goal, i.e.,* $s_\star \models g$.

The goal-based deontological principle allows pulling the lever, because the agent's goal is $s_\star = \mathit{men{=}alive}$ and $\mathit{men{=}alive}$ is morally good. Were it the case that the agent's goal actually was to bring about the death of the man on the other track, i.e., $s_\star = \mathit{man{=}dead}$, the goal-based deontological principle would judge pulling the lever morally impermissible, while the deontological principle according to Def. 1 would still permit pulling the lever. Pushing the man off the bridge in the footbridge dilemma is permitted by the goal-based deontological principle as long as the goal is to save the five people on the track but not when killing the big man was part of the agent's goal.

5.2 Consequentialism

Consequentialist ethical theories judge actions with respect to the consequences actions bring about. The most well-known consequentialist principle is utilitarian maximization. Besides this utilitarian principle, we also formalize principles that strictly forbid causal harm or instrumental harm.

5.2.1 Utilitarianism

The utilitarian principle requires an agent to always do what optimizes moral utility. In the context of action plans, we call a plan morally permissible accord-

ing to the utilitarian principle iff the final state of the plan is among the morally optimal states.

Definition 3 (Utilitarian Principle)**.** *Given a planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ *and an associated utility function* u*, a plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the utilitarian principle *if and only if* $u(s_n) \geq u(s')$ *for all states* s' *that are reachable given the set of actions* A*, where* s_n *is the final state reached by* π*.*

Given that the *advance* actions will be executed anyway, the set of reachable states in both the trolley problems boil down to the states reached by acting at time point 0 or by not acting at all. In the classical trolley problem, the two reachable states differ in the number of people dead. In our version of utilitarianism, the number of people harmed is morally relevant. Thus, the plan $\langle \mathit{pull} \rangle$ is morally permissible, but the empty plan is not. Likewise, for the footbridge trolley problem, pushing the big man off the bridge, $\langle \mathit{push} \rangle$, is morally permissible but the empty plan is not.

5.2.2 Harm Avoidance

While utilitarianism allows for harm for the greater good, it has been argued that autonomous agents should avoid to cause harm at all (Nevejans, 2016). We take a counterfactual approach to modeling harm by saying that an action causes harm if had the action not been performed, then the harm would not have happened.

Before presenting our definition of do-no-harm permissibility, we want to discuss a few simpler candidate definitions that might appear more obvious at first sight. However, as we will see, all of them have defects that necessitate the more complicated definition—which is also harder to verify computationally—to which we eventually commit ourselves.

The most straightforward candidate definition calls a plan do-no-harm permissible if nothing harmful is true in the final state. This is too strict, though, since the eventual harm may already be present in the initial state. For instance, whenever something harmful is true initially, the empty plan would be classified as impermissible, although it does not actively cause any harm. It just does not prevent it, either.

Repairing this, we might call a plan do-no-harm permissible if all harm that is true in the final state is already present initially, i. e., if no *additional* harm is done. This now turns out to be too weak, instead: if, for example, we have two actions in the plan, one deleting a harmful fact, which is true initially, and the second action reinstantiating the harm, then we are not worse off compared with the initial state. On the one hand, according to our repaired definition, this plan is do-no-harm permissible. On the other hand, we do not want to classify it as such, since in the intermediate state, avoidable harm is actively done.

We need to further improve the definition. The general idea is this: a plan is do-no-harm permissible if all harm that is potentially true in the final state is *unavoidable*. The crux is how to formalize (un-)avoidability. First, we can

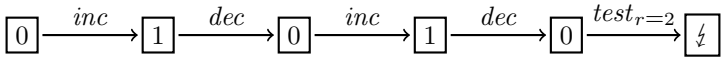
distinguish between avoiding harm by doing *less* and avoiding harm by doing *more*. Since the do-no-harm principle is about not actively causing harm, we only focus on avoiding harm by doing *less*. In other words, when judging do-no-harm permissibility, we will not explore alternative plans that actively *prevent* harm. Rather, we only consider alternative plans that passively *avoid* causing harm. Now, how much "doing less" do we need to consider?

The simplest option is to only consider avoiding harm by leaving out *one single action*. This is insufficient, however, since a plan may contain several actions that all make the same harmful fact true, a case of *overdetermination* (Lewis, 1973). Then, deleting neither of those actions alone would be sufficient to avoid the harm, and the plan would be classified as permissible. However, the harm could have been avoided by deleting all actions with the harmful effect, so it should not be classified as permissible.

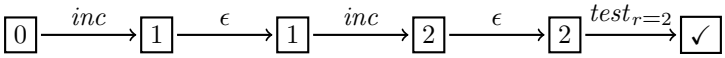
This observation suggests that we should consider eventual harm to be avoidable if it is avoidable by deleting some *subset of actions* in the plan. Again, it might be tempting to believe that only taking plan prefixes or suffixes into account for deletion is sufficient. To see that it is not, consider the following example: there is a resource r with possible values 0, 1, and 2, that can be incremented and decremented in steps of 1. Initially, $r{=}0$, and after a certain number of time steps, we need two units of the resource in order to stay away from harm. If we have fewer units, harm will be done (exogenously). More formally, $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, where:

$$
\begin{aligned}
\mathcal{V} &= \{r, h\}, \qquad \mathcal{D}_r = \{0,1,2\}, \qquad \mathcal{D}_h = \{\top, \bot\},\\
A &= A_{\text{endo}} \cup A_{\text{exo}}, \qquad A_{\text{endo}} = \{\mathit{inc}, \mathit{dec}, \epsilon\}, \quad A_{\text{exo}} = \{\mathit{test}_{r=2}\},\\
\mathit{inc} &= \langle \top, (r{=}0 \rhd r{:=}1) \land (r{=}1 \rhd r{:=}2) \rangle,\\
\mathit{dec} &= \langle \top, (r{=}2 \rhd r{:=}1) \land (r{=}1 \rhd r{:=}0) \rangle,\\
\mathit{test}_{r=2} &= \langle \top, (r{=}0 \rhd h{:=}\top) \land (r{=}1 \rhd h{:=}\top) \rangle, \qquad t(\mathit{test}_{r=2}) = \{4\},\\
s_0 &= r{=}0 \land h{=}\bot, \qquad s_\star = \top,\\
u(h{=}\bot) &= +1, \qquad u(h{=}\top) = -1, \quad \text{and}\\
u(r{=}d) &= 0 \quad \text{for all } d \in \{0,1,2\}.
\end{aligned}
$$

Plan $\pi = \langle \mathit{inc}, \mathit{dec}, \mathit{inc}, \mathit{dec} \rangle$ leads to harm, since after the four steps, $r{=}0$, and hence the exogenous action $\mathit{test}_{r=2}$ fails and produces harm:



However, it would have been possible to avoid harm by doing less, specifically by deleting the two instances of the *dec* action. For the subplan $\pi' = \langle \mathit{inc}, \epsilon, \mathit{inc}, \epsilon \rangle$, the test action succeeds and does not lead to harm.



Notice that the same result could not be achieved by only deleting a prefix or suffix of π. Also notice that this is not a particularly contrived example. In the real world, think of a first-responders scenario where two helpers are needed to ensure survival of an injured person. If a first responder arrives on the scene, then leaves, followed by a paramedic arriving (and possibly leaving afterwards), then the victim will die. Had the first responder (and possibly the paramedic) not left the scene, the victim would have survived. In this sense, the leaving actions were causal for the death of the victim, and the plan (arrive, leave, arrive, leave) is not permissible according to the do-no-harm principle.

In conclusion, we have to formalize causing of harm by refering to arbitrary subplans, which leads to combinatorial reasoning. One first shot at a formal notion of causality could look as follows:

A plan $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *causes* fact $v{=}d$ iff

1. $s_n \models v{=}d$, and
2. there exists a plan π', such that
   (a) π' results from replacing a subset of actions in π each by the empty action ϵ, and
   (b) in the final state $s'_{n'}$, the fact $v{=}d$ does not hold.

This formalization of causation does not suffer from the problem of overdetermination when only endogenous actions cause some fact. Consider a plan π involving two endogenous actions a_1 and a_2 both of which have $v{:=}d$ as an effect. Then, by deleting both a_1 and a_2 from π, $v{=}d$ is avoided. Things turn out differently when a_2 is an exogenous action. Consider $\pi = \langle a_1 \rangle$, and a_2 being executed after the first action. Under these circumstances, there is no way of avoiding $v{=}d$, therefore π does not count as a cause of $v{=}d$ even though it was sufficient for $v{=}d$ to become true. Intuitively, both the action plan and the exogenous actions are causes of $v{=}d$. We want to give precedence to the action plan. The following Definition 4 accomplishes this by allowing to first discard occurences of exogenous actions as long as their deletion does not make $v{=}d$ false in the final state, i.e., they were not essential for $v{=}d$.

Definition 4 (Causality)**.** *Given a planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$*, a plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ causes *fact* $v{=}d$ *iff*

1. $s_n \models v{=}d$,
2. *there exists a possibly empty subset of occurences of exogenous actions* O*, such that* $s_n \models v{=}d$ *still holds when the elements of* O *are deleted, and*
3. *the plan* π'*, which results from replacing a subset of actions in* π *each by the empty action* ϵ*, executed while discarding the occurences in* O*, leads to a final state* $s'_{n'}$ *in which* $v{=}d$ *does not hold.*

Defintion 4 can also handle cases of *preemption* (Lewis, 1973) which the first formalization could not. Consider the case of two shooters S1 and S2. Shooter S1

is the planning agent, that is, its shooting is modeled as an endogenous action. The second shooter, S2, is modeled as exogenous and shoots at time point 2 if and only if S1 does not shoot. In any case, the victim shot at will die. Consider the plan $\pi = \langle shoot \rangle$, which results in fact $dead{=}\top$. To show that π has caused $dead{=}\top$ according to our previous formalization , the plan $\pi' = \langle \epsilon \rangle$ must result in $dead{=}\top$ not to hold. However, because now S2 shoots, this is not the case. Definition 4 allows to delete the occurence of S2's shooting from the problem description, because doing so does not invalidate $dead{=}\top$ to become true after the execution of π. In this manipulated problem description, the plan π' avoids $dead{=}\top$, hence, π is a cause of $dead{=}\top$ according to Definition 4.

Based on this notion of causality, we now can define the *do-no-harm principle*.

Definition 5 (Do-No-Harm Principle)**.** *Given a planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ *and an associated utility function* u*, a plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the do-no-harm principle *if and only if* π *causes only neutral or good facts (according to* u*).*

According to this definition, the plan $\langle pull \rangle$ for the classical trolley problem is morally impermissible. This is because it makes the morally bad fact $man{=}deadOnTrack$ true, which is false if *pull* is deleted from the plan. For the analogous reason, the plan $\langle push \rangle$ for the footbridge trolley problem is impermissible, as well. Contrarily, the empty plan is permissible because the harm that results in the final state cannot be avoided by skipping actions.

5.2.3 Asimovian Principle

Definition 6 introduces the Asimovian principle, which seems to be more restrictive than do-no-harm. According to Asimov's first law of robotics, a robot should not cause harm and it should avoid harm to happen (Asimov, 1950). Hence, whereas the empty plan is always do-no-harm permissible, doing nothing is impermissible according to the Asimovian principle if there exists a plan which prevents the harm from holding in the final state. Particularly, in both the trolley problems, no morally permissible plan exists according to the Asimovian principle.

Definition 6 (Asimovian Principle)**.** *Given a planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ *and an associated utility function* u*, a plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the Asimovian principle *if and only if for all facts* $v{=}d$*, if* $s_n \models v{=}d$ *and* $u(v{=}d) < 0$*, then there is no alternative sequence of actions* π'*, such that* $s'_{n'} \not\models v{=}d$*, where* s_n *and* $s'_{n'}$ *are the final states reached by* π *and* π'*, respectively.*

Although the Asimovian principle looks like a stronger version of the do-no-harm principle, it is in fact incomparable. There are cases, which are permissibile according to the Asimovian principle, but which are not do-no-harm permissible. The reason for that is that in the Asimovian case, we take the exogenous actions as constant. So, in the case with the two shooters discussed above, the

plan $\langle shoot \rangle$ is Asimovian permissible, because there does not exist an alternative plan that prevents the death. However, it is not do-no-harm permissible according to Def. 5.

5.2.4 Instrumentality

A reasonable variation of the do-no-harm principle is the do-no-instrumental-harm principle. The idea is that harm is permissible in case it is not committed as a means to one's end but only occurs as side effect. In order to check for this, one would counterfactually ignore some subset of effects that assert the harmful fact $v_m{=}d_m$ in order to test whether this fact is a means to the goal $s_\star$. Similarly to the do-no-harm principle, we need to allow for the deletion of some other effects that are irrelevant for the goal before making the test in order to deal properly with cases of overdetermination and preemption. Furthermore, we would consider a harmful fact only as potentially instrumental if it is caused by the plan at all.

Definition 7 (Means to an end)**.** *Given a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ and a plan π with final state s_n, an assignment $v_m{:=}d_m$ is called a* means to the end $s_\star$*, if and only if*

1. $s_n \models s_\star$,
2. *there exists a possibly empty subset of assignments AS, such that after the deletion of the elements in AS from the effects of a subset of endogenous and exogenous actions, $s_\star$ still holds in the final state s_n of π, and*
3. *the additional deletion of $v_m{:=}d_m$ from the effects of a subset of actions in π leads to a final state in which $s_\star$ does not hold.*

Based on this definition, we can now define what it means that a plan is permissible according to the do-no-instrumental-harm principle.

Definition 8 (Do-No-Instrumental-Harm Principle)**.** *Given a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ and an associated utility function u, a plan $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ is* morally permissible according to the do-no-instrumental-harm principle *if and only if for all facts $v{=}d$, if $u(v{=}d) < 0$ and $v{=}d$ is caused by π (according to Def. 4), then $v{:=}d$ is not a means to an end.*

According to the definition of the do-no-instrumental-harm principle, the plan $\langle pull \rangle$ in the classical trolley dilemma is permissible. This is because the bad effect $man{:=}dead$ is not a means to the end $men{=}alive$. Contrarily, in the footbridge trolley problem, if, counterfactually, $man{:=}deadOnTrack$ was not an effect of *push*, the goal $men{=}alive$ would not finally hold. Hence, the plan $\langle pull \rangle$ is morally permissible according to the do-no-instrumental-harm principle, and $\langle push \rangle$ is not.

One is tempted to think that the do-no-instrumental-harm principle is at least as tolerant as the do-no-harm principle, i.e., every do-no-harm permissible plan is also do-no-instrumental-harm permissible. And this is indeed so because

of the definition. Only if a fact is caused by a plan, it will be considered as a culprit. So if no harmful fact is caused by the plan, and the plan is therefore do-no-harm permissible, then it must also be do-no-instrumental-harm permissible.

Proposition 1 (Instrumental Harm)**.** *Given a planning task* Π *and an associated utility function* u*, then any plan* π *that is do-no-harm permissible is also do-no-instrumental-harm permissible.*

5.3 The Principle of Double Effect

Finally, we define the principle of double effect, which uses many of the aforementioned principles.

Definition 9. *Given a planning task* $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ *and an associated utility funcition* u*, a plan* $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ *is* morally permissible according to the double-effect principle *if and only if all of the following conditions are satisfied:*

1. *The plan* π *is morally permissible according to the deontological principle.*
2. *At least one goal fact* $v{=}d$ *satisfies* $u(v{=}d) > 0$*.*
3. *No goal fact* $v{=}d$ *satisfies* $u(v{=}d) < 0$*.*
4. *The plan* π *is morally permissible according to the do-no-instrumental-harm principle.*
5. $u(s_n) > 0$*, where* s_n *is the goal state reached by* π*.*

Hence, the principle of double effect contains the deontological principle as its first condition and the do-no-instrumental-harm principle as the fourth condition. The second and third conditions are constraints on the goal of the planning agent: She is not allowed to have morally bad goals, and the goal should contain something morally good. The last condition is a weaker form of utilitarianism, which requires that all in all the plan brings about more good facts than bad facts—but unlike utilitarianism, it does not require the plan's final state to be among the optimal states.

In case of the footbridge trolley problem, the first condition renders pushing the man off the bridge impermissible. However, the second and third conditions are fulfilled, because the goal of the agent only consists of one fact, viz., *men=alive*, and this fact is morally good. The fourth condition also is violated as we have already discussed above. The fifth condition is fulfilled, because, all in all, the good consequences yield more positive utility than the negative consequence add negative utility. Hence, using the principle of double effect, the reasoner can explain that there are two reasons why the plan $\langle push \rangle$ is morally impermissible: Because pushing is morally bad, and because the death of the big man is used as a means. For the case of the classical trolley problem, the principle of double effect comes to the conclusion that the plan $\langle pull \rangle$ is morally permissible.

6 A Note on Modeling

It is often the case that a planning domain could be modelled in different ways, which may lead to equivalent results. Sometimes, it is the case that these different ways impose different search costs, but nevertheless one solves the same problem. If we add the ethical dimension, however, results can significantly differ based on what kind of domain modeling we choose. In order to illustrate this, let us consider an example, where two persons are drowning in two different lakes. The planning agent can walk to the first lake and rescue the first person. Instead of rescuing the first person, the agent could walk on and rescue the second person. At time point 3, all persons still in the lake will drown. This could be modelled as follows. The variable l captures the location of the agent (0 = start point, 1 = lake 1, 2 = lake 2), p_i captures whether person i is alive and s_i captures whether person i is swimming:

$$
\begin{aligned}
\mathcal{V} &= \{l, p_1, p_2, s_1, s_2\}, \qquad \mathcal{D}_l = \{0,1,2\}, \qquad \mathcal{D}_{p_i} = D_{s_i} = \{\top, \bot\},\\
A &= A_{\text{endo}} \cup A_{\text{exo}}, \qquad A_{\text{endo}} = \{\mathit{walk}, \mathit{rescue}, \epsilon\}, \quad A_{\text{exo}} = \{\mathit{drown}\},\\
\mathit{walk} &= \langle \top, (l{=}0 \rhd l{:=}1) \wedge (l{=}1 \rhd l{:=}2)\rangle,\\
\mathit{rescue} &= \langle \top, (l{=}1 \rhd s_1{:=}\bot) \wedge (l{=}2 \rhd s_2{:=}\bot)\rangle,\\
\mathit{drown} &= \langle \top, (s_1{=}\top \rhd p_1{:=}\bot) \wedge (s_2{=}\top \rhd p_2{:=}\bot)\rangle, \qquad t(\mathit{drown}) = \{3\},\\
s_0 &= l{=}0 \wedge p_1{=}\top \wedge p_2{=}\top \wedge s_1{=}\top \wedge s_2{=}\top,\\
s_\star &= \top,\\
u(p_i{=}\bot) &= -1,\ u(p_i{=}\top) = 1 \qquad \text{and} \qquad u(v{=}d) = 0 \text{ for all } v \in \{l, s_1, s_2\}.
\end{aligned}
$$

Let us now consider the plan $\pi = \langle \mathit{walk}, \mathit{walk}, \mathit{rescue}\rangle$. This plan produces the harmful effect $p_1{=}\bot$. If we now consider the plan $\pi' = \langle \mathit{walk}, \epsilon, \mathit{rescue}\rangle$, we see that this plan does not produce this harmful effect, i.e., π causes $p_1{=}\bot$ and therefore plan π is not do-no-harm permissible. On the other hand, π' is do-no-harm permissible because no deletion of any endogenous action can save person 2. Thus the ethical principle seems to prefer to save the first person, which does appear to be very counter-intuitive.

However, this preference appears to be a modeling artifact. While rescuing a person in lake 1 or 2 is the same *type* of action, these two actions are different action *tokens* in the sense that the parameters and the context of the execution action are different. In fact, in a planning language with schema variables (such as PDDL), one would have an operator *rescue* with at least two parameters, namely, location and person. So, in the above formalization, one should have two different actions rescue_1 and rescue_2 for rescuing a person in lake 1 and lake 2, respectively. Similarly, one should have two different walk actions. This would mean that π should have the form $\langle \mathit{walk}_1, \mathit{walk}_2, \mathit{rescue}_2\rangle$. Replacing now either walk_1 and walk_2 by ϵ would not help person 1. So, the plan would be do-no-harm permissible.

This example demonstrates that the modeling can have a crucial influence on the ethical verification of plans and should be done carefully. Another significant influence is introduced by how the utility function u is defined. As the approach

to ethical reasoning dealt with in this article focuses on reasoning about what is morally right (action sequences) as opposed to what is morally good (individual actions and facts), we assume u to be externally given, either by the modeler or by some external process.

Some ethical situations may require to assign different utility values to similar action types due to different contexts. For instance, a modeler may want to distinguish killing in self-defense from killing per se. To make this distinction, one can add two different endogenous actions: Action *kill-in-self-defense* has precondition that the agent is currently under attack, whereas for the action *kill* this precondition is missing. Then, *kill-in-self-defense* and *kill* can be assigned different utility values. A related problem concerns action order: getting married first and than having children may be judged differently from having children first and then getting married. Again, one solution consists in introducing two actions *have-children1* and *have-children2*, such that the first action has *married* among its preconditions while the second has not. Then these two actions can again receive different utility values.

7 Ethical Evaluation of Action Plans and Computational Complexity

The output of a planning algorithm is a sequence of actions $\pi = \langle a_0, \ldots, a_{n-1} \rangle$ and a final state s_n. Our goal is to ethically evaluate a given action plan. To this end, we here describe procedures that for each ethical principle take a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, the utility function u, a plan π, and its final state s_n, as the input and decide whether or not the principle renders the plan as morally permissible. Furthermore, we determine the computational complexity of the evaluation problem for the different principles, which turn out to be surprisingly hard (see Table 1). The ethical evaluation problem can be stated as follows.

Definition 10 (Ethical evaluation problem relative to ethical principle $\mathcal{E}$)**. Given:** *A planing task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, a utility function u, and a plan π that transforms s_0 into a state satisfying $s_\star$.*

Asked: *Is the plan π morally permissible according to ethical principle $\mathcal{E}$?*

To check whether or not a given plan π is morally permissible according to the deontic principle (Def. 1), it needs to be checked if some of the actions in π are intrinsically bad, i.e., if for one of the actions a_i in π, we have $u(a_i) < 0$. This can be apparently done in time linear in the length of Π and π. The goal-based deontological principle (Def. 2) can also be checked in linear time by checking that none of the conjuncts in the goal specification $s_\star$ has negative utility.

Proposition 2 (Deontic Evaluation)**.** *Deciding whether a plan is morally permissible according to the deontological principles can be done in polynomial time.*

A procedure for verifying that π is morally permissible according to the utilitarian principle (Def. 3) is much more involved than checking deontological permissibility. Recall that the utilitarian principle only permits plans that

Ethical principle	Computational complexity
Act-/Goal-based Deontology	linear time
Utilitarianism	PSPACE-complete
Do-no-harm principle	co-NP-complete
Asimovian principle	PSPACE-complete
Do-no-instrumental-harm principle	co-NP-complete
Doctrine of double effect	co-NP-complete

Table 1: Computational complexity of the ethical evaluation problem

lead to reachable states with maximum utility. In so far, this is very similar to over-subscription planning (Smith, 2004). Based on that, we can formulate a non-deterministic procedure for deciding the complement of the permissibility problem as follows: Compute the overall utility of s_n. Then guess another complete state s' with utility that is larger than the utility of s_n. Finally generate (non-deterministically) a plan π' to achieve s'. If successful, it demonstrates that π is not morally permissible. That this is indeed an (asymptotically) optimal procedure is shown by the following theorem.

Theorem 1 (Utilitarian Evaluation)**.** *Deciding whether a plan is morally permissible according to the utilitarian principle is PSPACE-complete.*

Proof. PSPACE membership follows from the arguments above, and the facts that PSPACE is closed under complement and non-determinism and that deciding plan existence (in SAS^+) is in PSPACE. PSPACE-hardness follows straightforwardly from a reduction of plan existence in SAS^+ planning. Given a SAS^+ planning task Π, generate a new task Π' by extending the set of variables by two Boolean variables g_1 and g_2, which are both assumed to be false in s_0. Extend the set of actions by two new endogenous actions: $a_1 = \langle \top, g_1{:=}\top \rangle$ and $a_2 = \langle s_\star, g_2{:=}\top \rangle$. The new goal description of Π' is $s_\star = g_1{=}\top$. The utility function is identical to zero on all actions and facts except for g_1 and g_2, where it evaluates to 1. Clearly, one possible plan is $\langle a_1 \rangle$ leading to state s with $u(s) = 1$. This plan is impermissible according to the utilitarian principle iff there exists a plan for the original task Π because in this case we could reach a state s' for Π' such that $u(s') = 2$. □

To check whether a given plan π is morally permissible according to the do-no-harm principle (Def. 5), we have to verify that no parts of the plan lead to avoidable harm after we may have removed a subset of the exogenous action occurrences that are not relevant for the harmful fact. A non-deterministic algorithm for deciding impermissibility could be: We guess one fact $v_b{=}d_b$ with $u(v_b{=}d_b) < 0$, a subset of exogenous action occurrences O, and a sub-plan π' of π, where some actions are replaced by empty actions. We then need to verify the three conditions mentioned in Def. 4. First, we have to verify $s_n \models v_b{=}d_b$. Then we verify that the plan π still produces $v_b{=}d_b$, even if all exogenous action

occurrences from O are deleted. Finally, we need to verify that $v_b{=}d_b$ is not produced by π' under the condition that the O actions are deleted.

Theorem 2 (Do-No-Harm Evaluation)**.** *Deciding whether a plan is morally permissible according to the do-no-harm principle is co-NP-complete.*

Proof. The sketched non-deterministic algorithm demonstrates membership of impermissibility in NP, hence permissibility is in co-NP. In order to show hardness, we use a reduction from 3SAT to the impermissibility problem. Assume a 3SAT problem over the variables $v_1, \ldots, v_n$ and clauses $c_1, \ldots, c_m$, where each clause consists of 3 literals l_{j1}, l_{j2}, l_{j3}. We now construct a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$, where

$$\begin{array}{rcl} \mathcal{V} & = & \{b, g, v_1, \ldots, v_n, c_1, \ldots, c_m\}, \\ A & = & \{V_1, \ldots, V_n, C_1, \ldots, C_m, G, B\}, \\ s_0 & = & \{v{=}\bot \mid v \in \mathcal{V}\}, \text{and} \\ s_\star & = & \{g\}. \end{array}$$

The actions are defined as follows: $V_i = \langle \top, v_i{:=}\top \rangle$, $C_j = \langle \top, \bigwedge_{k=1}^{3}(l_{jk} \rhd c_j) \rangle$, where $l_{jk} \equiv v_{jk}{=}\top$ if the literal l_{jk} in the original SAT problem is positive, otherwise, $l_{jk} \equiv v_{jk}{=}\bot$. Further, $G = \langle \top, g{:=}\top \wedge (\bigwedge_{j=1}^{m} c_j \rhd b{:=}\top) \rangle$, $B = \langle \top, b{:=}\bot \rangle$, and we assume that there are no exogenous actions. All facts have zero utility except for $b{=}\bot$, which is valued -1. The plan we want to check is $\pi = \langle V_1, \ldots, V_n, C_1, \ldots, C_m, G, B \rangle$. This plan obviously achieves the goal and the final state contains some harm. Moreover, the only way to avoid this harm is to delete action B. However, even without this action, we still may have harm. This harm can be avoided, if and only if we can delete a (perhaps empty) subset of the V_i actions corresponding to a variable assignment of the 3SAT problems that satisfies the original 3SAT formula, which demonstrates that impermissibility is NP-hard, i.e., permissibility is co-NP-hard. □

For the Asimovian principle, for each harm in the final state, we have to check whether there is a plan, which avoids that harm. As for the utilitarian principle, this quantifies over all available plans, and hence checking Asimovian permissibility has the same computational complexity as utilitarian permissibility.

Theorem 3 (Asimovian Evaluation)**.** *Deciding whether a plan is morally permissible according to the Asimovian principle is PSPACE-complete.*

Proof. We first prove membership by presenting a procedure which uses polynomial space: Consider as input a planning task $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ and a plan π. As a first step, the execution of π is simulated to obtain the final state s_n, whose size is bound by $|\mathcal{V}|$. For all harmful facts $v{=}d$ (viz., with $u(v{=}d) < 0$) that hold in s_n, a planner is used to solve the planning tasks $\Pi_{d_i} = \langle \mathcal{V}, A, s_0, v{=}d_i \rangle$ for each $d_i \in \mathcal{D}_v$ with $d_i \neq d$. If for one of the d_i's such a plan is found, this demonstrates that $v{=}d$ can be avoided, i.e., that the original plan is not Asimovian permissible. Plan existence is known to be decidable in polynomial space, i.e.,

with that it follows that impermissibility, and therefore also permissibility, is in PSPACE.

To show hardness, we reduce SAS^+ plan existence to Asimovian permissibility: Let $\Pi = \langle \mathcal{V}, A, s_0, s_\star \rangle$ be a SAS^+ planning instance. Now create a new instance $\Pi' = \langle \mathcal{V} \cup \{g\}, A \cup \{G\}, s_0 \wedge g{=}\bot, g{=}\top \rangle$, where g is a new variable with $\mathcal{D}_g = \{\top, \bot\}$ and $G = \langle s_\star, g{:=}\top \rangle$ is a new action. Set $u(g{=}\bot) = -1$ and $u(f) = 0$ for all other facts f. The empty plan $\pi_\epsilon = \langle\rangle$ is morally impermissible according to the Asimovian principle iff there exists a plan that solves Π' (avoiding the harm $g{=}\bot$). □

For the do-no-instrumental-harm principle (Def. 8), we can use a very similar method to checking for the do-no-harm principle. Instead of skipping subsets of actions, we have to delete subsets of effect occurrences in the plan. Hence, checking this principle for a given plan has the same computational complexity.

Theorem 4 (Do-No-Instrumental-Harm Evaluation)**.** *Deciding whether a plan is morally permissible according to the do-not-instrumental-harm principle is co-NP-complete.*

Proof. One can use a similar non-deterministic algorithm as for the do-no-harm principle, demonstrating that deciding permissibility of a plan for this principle is again in co-NP. For hardness, we can use a reduction very similar to the one in Theorem 2. Instead of deleting actions we would delete effects, which are used to enable the execution of exogenous actions that regulate the assignment of the variables. □

Finally, we consider the double-effect principle. Except for the fourth condition, everything can be checked in polynomial time. The fourth condition is just the do-not-instrumental-harm principle. In other words, deciding permissibility for this principle is again in co-NP.

Theorem 5 (Double Effect Evaluation)**.** *Deciding whether a plan is morally permissible according to the double-effect principle is co-NP-complete.*

Proof. Membership is obvious. Hardness follows with a similar proof as above. □

8 Related Work

While there exists a number of papers on machine ethics, papers that focus on generating and/or validating plans according to ethical principles are scarce. A general survey about different etethical machine reasoning approaches has been recently published by Dennis and Fisher(2018).

Dennis et al. (2016) propose to establish ethical principles and ethical rules that judge the severity of violating ethical principles, whereby an ethical principle could be not to harm a human. Plans can then be ordered by comparing the worst violations of these plans. While this has an deontological flavor, in

fact, plans are judged according to their ultimate consequences, and hence this appears to be a consequentialist approach. The authors do not consider the distinction between causing harm and causing instrumental harm.

Pereira and Saptawijaya (2017) use abductive logic programming in order to specify the principle of double effect and to evaluate some of the trolley scenarios. Berreby et al. (2015) similarly use logic programming (in this case ASP) in order to specify the principle of double effect and to evaluate this formalization on trolley scenarios described using the event calculus. In this case, however, they do not use counterfactual reasoning to judge causality, but they use simple syntactical means to determine what is a cause of an effect. Govindarajulu and Bringsjord (2017) propose a general framework to create or verify that an autonomous system is compliant to the double effect principle. For this purpose they introduce a powerful logical formalism called *deontic cognitive event calculus*. In addition, they propose a formalization of the notion of *means to an end* in a STRIPS framework, which however does not take into account that different actions in a plan can contribute to different parts of a goal, and which does not consider that combinations of actions can be causes. Weld and Etzioni (1994) propose two versions of a do-no-harm principle for action plans. Their do-no-harm principle is fine with harm in the final state given that the harm already held in the initial state. Our do-no-harm principle does not permit to heal harm first just to reintroduce it later on. However, our formulation allows to cause harm if it is healed later on. This is not allowed in one version of Weld and Entzioni's account. We can, however, generate this behavior by introducing special harm facts into the model that become true when harm happens during plan execution and that remain true forever.

If we move from ethical reasoning to reasoning about norms in general, then there is a large body of work to consider. One example is the survey paper on monitoring norms by Dastani et al. (2018). They throw a very wide net covering legal, social, moral, and rational norms. When it comes to the specification of norms, they mention LTL, finite state autmomata, Petri nets, and more. However, they do not mention counterfactual analysis of plans. Indeed, one interesting question is how one could express the ethical principles we considered using LTL or similar logics. While it is hard to rule the impossibility in general, complexity results of LTL on finite traces (Fionda & Greco, 2016) seem to suggest that for LTL an exponential blowup of the LTL specification is to be expected.

None of the papers mentioned above address the issue that evaluating the moral permissibility of action plans might require a counterfactual analysis that is combinatorial in nature. However, in the context of causal explanations of plan failures, Alechina and colleagues (Alechina, Halpern, & Logan, 2017) and Goebelbecker and colleagues (Göbelbecker, Keller, Eyerich, Brenner, & Nebel, 2010) present complexity results. Different to our work, these authors do not analyze moral principles, they do not consider exogenous events, but instead they care about reasons for plan failure while we are interested in deciding if single facts were caused by a plan or used as a means.

9 Conclusions and Outlook

We formalized various ethical principles, which consider different aspects of a plan to be morally significant. Deontology stresses the moral value of action tokens, utilitarianism requires utility optimization, the do-no-harm principle and the Asimovian principles strive for avoiding avoidable harm, and the do-no-instrumental-harm principle and the principle of double effect take seriously the intuition that harm should not be used as a means to an agent's end but may be acceptable as a mere side effect.

We studied these ethical principles in the context of *action sequences*, as opposed *individual actions*. Only in this way we can analyze moral permissibility of entire plans, since it is not sufficient to judge the moral permissibility of each action in isolation, but also in the context of the whole plan. Overdetermination and preemption caused complications in the ethical reasoning process that led to a jump in computational complexity. We showed that, with respect to our formalization, verification is PSPACE-complete for utilitarianism and for the Asimovian principle, co-NP-complete for do-no-harm, for do-no-instrumental-harm, and for the principle of double effect, and that it is polynomial-time for deontology. Verifying the do-no-harm principles involves a combinatorial reasoning over possible *sets* of actions that lead to harm or that may be instrumental towards achieving a goal condition, which makes verifying those ethical principles surprisingly hard.

We believe that our work has the potential of being useful in making autonomous systems ethical by providing them with the capability of coming up with morally permissible plans or at least being able to judge ethical permissibility of given plans.

Future work includes covering algorithmic aspects of moral *planning*, i. e., how to come up with morally permissible plans efficiently, as opposed to mere moral *plan evaluation*. This specifically means describing and empirically evaluating different moral planning algorithms. Furthermore, the framework proposed in this manuscript takes the goal to be reached for granted. In cases where such a goal cannot be achieved in a morally permissible way, it may be called for to revise the goal such that the task becomes solvable by a permissible plan, while retaining as much of the original objective as possible.

Another interesting avenue of research is to generalize our approach to other forms of planning, such as multi-agent epistemic (Engesser, Bolander, Mattmüller, & Nebel, 2017) and/or probabilistic planning (Goldman & Zilberstein, 2004).

References

Alechina, N., Halpern, J. Y., & Logan, B. (2017). Causality, responsibility and blame in team plans. In *AAMAS '17 Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (pp. 1091–1099).

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge, UK: Cambridge University Press.

Anderson, M., Anderson, S. L., & Armen, C. (2005). *Machine Ethics: Papers from the AAAI Fall Symposium* (Tech. Rep.). AAAI Press.

Asimov, I. (1950). I, Robot. In *I, Robot* (chap. Runaround). Gnome Press.

Bäckström, C., & Nebel, B. (1995). Complexity results for SAS^+ planning. *Computational Intelligence*, *11*(4), 625–655.

Bentzen, M. M. (2016). The double effect principle applied to ethical dilemmas of social robots. In *What Social Robots Can and Should Do* (pp. 268–279).

Berreby, F., Bourgne, G., & Ganascia, J. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *Proceedings of the 20th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 2015)* (pp. 532–548).

Cresswell, S. N., & Coddington, A. M. (2003). Planning with timed literals and deadlines. In *Proceedings of the 21st Workshop of the UK Planning and Scheduling SIG* (p. 22-35).

Dastani, M., Torroni, P., & Yorke-Smith, N. (2018). Monitoring norms: a multi-disciplinary perspective. *Knowledge Eng. Review*, *33*, e25. Retrieved from `https://doi.org/10.1017/S0269888918000267` doi: 10.1017/S0269888918000267

Dennis, L. A., & Fisher, M. (2018). Practical challenges in explicit ethical machine reasoning. In *International Symposium on Artificial Intelligence and Mathematics, (ISAIM-2018)*. Retrieved from `http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Ethics_Dennis_Fischer.pdf`

Dennis, L. A., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, *77*, 1–14.

Driver, J. (2006). *Ethics: The Fundamentals*. Hoboken, NJ: Wiley-Blackwell.

Engesser, T., Bolander, T., Mattmüller, R., & Nebel, B. (2017). Cooperative epistemic multi-agent planning for implicit coordination. In *Proceedings of the Ninth Workshop on Methods for Modalities (M4M@ICLA-2017)* (pp. 75–90). Retrieved from `https://doi.org/10.4204/EPTCS.243.6` doi: 10.4204/EPTCS.243.6

Fionda, V., & Greco, G. (2016). The complexity of LTL on finite traces: Hard and easy fragments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*. Retrieved from `http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12250`

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5-15.

Fox, M., Howey, R., & Long, D. (2005). Validating plans in the context of processes and exogenous events. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)* (pp. 1151–1156).

Ghallab, M., Nau, D. S., & Traverso, P. (2016). *Automated Planning and Acting*. Cambridge University Press.

Göbelbecker, M., Keller, T., Eyerich, P., Brenner, M., & Nebel, B. (2010). Coming up with good excuses: What to do when no plan can be found. In

Proceedings of the 20th International Conference on Automated Planning and Scheduling (ICAPS 2010) (pp. 81–88).

Goldman, C. V., & Zilberstein, S. (2004). Decentralized control of cooperative systems: Categorization and complexity analysis. *J. Artif. Intell. Res.*, *22*, 143–174. Retrieved from `https://doi.org/10.1613/jair.1427` doi: 10.1613/jair.1427

Govindarajulu, N., Bringsjord, S., Ghosh, R., & Sarathy, V. (2019). Toward the engineering of virtuous machines. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society (AIES) 2019.*

Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)* (pp. 4722–4730).

Helmert, M. (2006). The fast downward planning system. *Journal of Artificial Intelligence Research*, *26*(1), 191–246.

Hölldobler, S. (2018). Ethical decision making under the weak completion semantics. In *Proceedings of the Workshop on Bridging the Gap between Human and Automated Reasoning* (pp. 1–5).

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lindner, F., & Bentzen, M. M. (2018). A formalization of kant's second formulation of the categorical imperative. In *Proceedings of the 14th International Conference on Deontic Logic and Normative Systems (DEON).*

Lindner, F., Bentzen, M. M., & Nebel, B. (2017). The HERA approach to morally competent robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017)* (pp. 6991–6997).

Mangan, J. (1949). An historical analysis of the principle of double effect. *Theological Studies*, *10*, 41–61.

Nevejans, N. (2016). *European Civil Law Rules in Robotics.* European Union.

Pereira, L. M., & Saptawijaya, A. (2017). Agent morality via counterfactuals in logic programming. In *Proceedings of the CogSci 2017 Workshop on Bridging the Gap between Human and Automated Reasoning—Is Logic and Automated Reasoning a Foundation for Human Reasoning?* (pp. 39–53).

Rintanen, J. (2003). Expressive equivalence of formalisms for planning with sensing. In *Proceedings of the 13th International Conference on Automated Planning and Scheduling (ICAPS 2003)* (pp. 185–194).

Smith, D. E. (2004). Choosing objectives in over-subscription planning. In *Proceedings of the 14th International Conference on Automated Planning and Scheduling (ICAPS 2004)* (pp. 393–401).

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Weld, D. S., & Etzioni, O. (1994). The first law of robotics (A call to arms). In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI 1994)* (pp. 1042–1047).